Journal of Construction Education
Spring 2000, Vol. 5, No. 1, pp. 20-29

1522-8150/00/$3.00/Educational Practice Manuscript

Student Teaching Evaluations: Options and Concerns

Bolivar A. Senior
Colorado State University
Fort Collins, Colorado

Student teaching evaluations attempt to provide information about an instructor's teaching effectiveness. While the majority of evaluations consist of written questionnaires administered at the end of the academic term, mid-term evaluations can provide timely and useful feedback. Many instructors regard the use of evaluations for personnel decisions with distrust, especially since concerns have been raised about the validity and bias of these instruments. Although many articles have been written about teaching evaluations, this topic still needs to be researched and discussed in construction management education. This article discusses the most common approaches to student teaching evaluation, as well as the concerns that have been raised by educators and researchers about their use. Further research needs are discussed in its Conclusion.

Keywords: Teaching evaluation, Instructional improvement, Grading leniency, Dr. Fox effect.

Introduction

The vast majority of American higher education institutions require some form of student evaluation of teaching (Newport, 1996; Wachtel, 1994). These evaluations have substantial influence in administrative decisions such as faculty tenure, promotion, and post-tenure reviews. On the other hand, teaching evaluations provide feedback to instructors about their pedagogical strengths and weaknesses (Marsh and Roche, 1997). With such important issues at stake, literally thousands of articles have been written about student evaluations (Marsh and Dunkin, 1992). This vast amount of information has not yielded a consensus about how they should be used, or even whether they should be used at all (Wachtel, 1998). Opposing views are manifested clearly and loudly in the literature. For example, a full section of *American Psychologist* (Vol. 52 (11), November 1997) devoted to the topic had rebuttals and counter-rebuttals of the very articles presented on the issue.

The ongoing, passionate debate over student evaluations has not been apparent in construction management education literature. A review of the ASC Conference Proceedings reveals articles on related issues such as faculty tenure and promotion (e.g., Christensen et al., 1992, Ciesielski, 1997), outcomes assessment (e.g., Segner and Arlan, 1991, Slobojan 1992, Yoakum, 1994, Hauck, 1998), and student peer evaluations (Feigenbaum and Holland, 1997), but no article has had a primary focus on teaching evaluations. This article discusses the most common approaches to student teaching evaluations, as well as the concerns that have been raised by educators and researchers about their use. To provide a context within construction management education, examples of teaching evaluation items and a mid-term evaluation session at Colorado State University are included here.

Written Questionnaires: Basic Building Blocks of Student Evaluations

Student questionnaires are, by far, the most popular form of teaching feedback (Simpson, 1995). These paper-and-pencil tools are easy to tabulate and economical to administer, but they are also the most controversial evaluation tool, due to common problems in their design and implementation.

One of the most quoted questionnaire designs is the Student's Evaluation of Educational Quality (SEEQ), developed by Marsh et al. (Marsh, 1987, Marsh and Dunkin, 1992). It identifies nine factors that should be probed: Learning / Value, Instructor Enthusiasm, Organization / Clarity, Group Interaction, Individual Rapport, Breadth of coverage, Examinations / Grading, Assignments / Readings, and Workload / Difficulty. The number and phrasing of questions is left open to the needs of each college, but since questions are often drafted by non-specialists in pedagogical issues (Newport, 1996), poorly structured items can slip into a questionnaire. A recent questionnaire in a mid-western university asked construction management students to rate from "Strongly Agree" to "Strongly Disagree" the statement "Overall, I would rate this instructor as excellent." If the instructor were good, but not excellent, how would a student answer this question? The author contacted the university and the actual drafter of the question. She explained that the original question was "How would you rate this instructor," with answers ranging from "Excellent" to "Poor." The magic of an administrative committee created its final published form. The drafting of teaching evaluation questionnaires by administrative committees has been frequently criticized in the literature (e.g., Lin et al., 1984; McKeachie, 1997), since the resulting material tends to emphasize summative applications.

Newport (1996) recommends limiting student questionnaires to low-inference items such as "The instructor began class on time," "The course syllabus included the course objectives," and "The instructor was prompt in returning tests and written assignments." Newport also advocates the elimination of high-inference items such as "The instructor was skilled at observing student reactions and modified his instructional strategies when needed" and "The instructor served as a good model of a reflective decision-maker." While low-inference items refer to facts or behavior readily observable by students, high-inference items require more sophisticated judgment.

High-inference items are ubiquitous in the questionnaires reviewed by the author. At Colorado State University, for instance, one mandatory survey includes the questions "Course assignments are consistent with the objectives of the course", and "The assessments / assignments / examinations were appropriate and clear" (CSU Student Course Survey). Another mandatory questionnaire at CSU asks students if "The grading system was fair", "The instructor utilized a variety of teaching styles", and "The instructor created an atmosphere conducive to learning." (Teaching Evaluation Instrument, CAHS). These questionnaires were introduced on fall 1998. A tabulation of answers to these items was not available for analysis when this article was written.

Mid-term Evaluations: Valuable Alternatives

A mid-term evaluation can consist of only a conventional student questionnaire, but this feedback technique is very commonly augmented with interviews, debriefing sessions and follow-up questionnaires. From a pedagogical perspective, there is evidence that mid-term evaluations are substantial improvements over end-of-term questionnaires. Studies by Abbott et al. (1990) and Wulff et al. (1985) showed that students are more satisfied with mid-term evaluations. Furthermore, Cohen (1980) concluded, "Instructors who received midterm feedback were subsequently rated about one-third of a standard deviation higher than controls." Even with such positive comments, most faculty do not use mid-term evaluations. Jacobs (1987) found that 82% of instructors only use end-of-term evaluations, and that 28% of them administered the evaluation on the last day of class.

The most common form of mid-term evaluation is the Small Group Instructional Diagnosis (SGID). It is an open-ended technique developed at the University of Washington's Biology Learning Resource Center from a model created by Melnik and Allen, University of Massachusetts (Clark and Bekey 1979). The essential procedure for SGID is described in the following section. Many other mid-term feedback instruments have been developed. For example, Fabry et al. (1997) describe a series of continuous feedback instruments that were administered regularly during a semester. The most effective (for formative purposes) and popular (among students) of these instruments was The Muddiest Point, where students wrote, anonymously, the most unclear point at the end of each lecture. Fabry et al. consider that because grades are still unsettled, student anonymity is essential for mid-term evaluations. In contrast, Timpson and Bendel-Simso (1996) encourage the full and public participation of students in the feedback process.

Effectiveness of Mid-term Evaluations: An Example

The author was involved in a difficult situation where a mid-term evaluation was of great help. He was co-teaching a capstone project management course at Colorado State University, and there was a subtle but unmistakable negative environment in the class. On his request, the Center for Teaching and Learning at CSU conducted a mid-semester student feedback session that essentially followed the SGID model, but included elements developed at CSU. The session was completed in less than one hour of regular meeting time. The facilitator explained to the students how the meeting would be conducted, and that it had been voluntarily requested by the instructor. It consisted of three parts. First, the instructor left the classroom, and the conventional questionnaire form used at CSU for teaching evaluations was administered. Each student also included three positive comments about the instructor, and the three main concerns about the course. The second part was also conducted without the instructor's presence, and consisted of a discussion of the students' concerns, including recommendations to the instructor. In the third part, the instructor returned to the classroom and was debriefed about the concerns and recommendations.

It turned out that the students did not question the instructor's level of knowledge, and consistently appreciated his sensitivity towards them as individuals. The main concerns were a

perceived disorganization and a tendency to stray to unrelated subjects during the lecture (both issues coming as total surprises to the instructor). It was relatively simple to improve the manifested concerns, and the end-of-term evaluations were only slightly below the instructor's average.

As a result of this experience, the author has continued conducting a modified mid-term evaluation in other courses. A questionnaire is administered after the first six weeks of the semester, and then repeated in the twelfth week. The form used contains the same questions used in the university-wide mandatory survey, and also requests three positive comments about the course and / or the instructor, three concerns about the course, and open-ended comments. The survey is administered at the beginning of the class period, and takes fifteen minutes to complete. The results are discussed at the end of the next session, and students provide specific recommendations to the instructor. In one case, a follow-up questionnaire was used to preserve anonymity, but students have been very willing to voice any negative issues despite the lack of anonymity in this part of the process.

Concerns about Student Evaluations: A Serious Issue

Despite all the benefits that student feedback can bring to the classroom (Marsh, 1987), the use of student evaluations for instructional improvement is dismally infrequent. A survey by Spencer and Flyr (1992) found that only 23% of faculty made changes to their teaching based on student evaluations, and that the majority of these changes were superficial, such as altering handouts, modifying presentation habits, and changing assignments. Reasons for the infrequent use of student evaluations to improve teaching can probably be traced to the concerns that many instructors harbor about the fairness and usefulness of these surveys. Ryan et al. (1980), found that the mandatory use of student evaluations led "to a significant reduction in faculty morale and job satisfaction." They also reported cases where instructors lowered standards and workloads, developed easier examinations, and probably inflated students' grades. Baxter (1991) found that in cases where evaluations are left to the instructor's discretion, such negative impacts were much lower. The following sections explore some of the most common concerns about student evaluations raised in the literature, without claiming to be a comprehensive review of such concerns. Extensive reviews have been written by Wachtel (1998) and d'Apollonia and Abrami (1997), among others.

Formative and Summative Applications: at the Heart of Concerns

Formative applications are those where student evaluations are used as a tool for instructional improvement. In contrast, summative applications make use of evaluations for administrative purposes such as decisions about faculty retention, tenure, promotion and salary increases.

The use of ratings for personnel decisions has been criticized by many authors (e.g., Murray, 1984, Ramsden and Dodds, 1989). The worst scenario comes when instructors are ranked from "best" to "worst" based on their student ratings. D'Apollonia and Abrami (1997) point out that this approach implies that 50% of the faculty fall "below the norm," even if they are excellent

teachers. Imagine a baseball team where all players are batting over .300. Ranking all players from best to worst would imply that 50% have averages "below the norm," even though the worst player would be quite competent The opposite case could also happen. A team's "best" player could average .250, which by most standards is low. As McKeachie (1997) points out, small numerical differences are "unlikely to distinguish between competent and incompetent teachers."

The antagonism to the summative use of student evaluations has resulted in caustic articles and commentaries. For example, Newport (1996) writes that:

> "Few of the higher education administrators in the USA who rely on amateur raters to assess teaching performance [...], would allow untrained and inexperienced students to cut their hair, [...] or to make investment decisions involving a few thousand dollars of their personal funds. [...] Yet, in the USA, untrained, amateur student raters are routinely used in making salary adjustments, tenure and promotion decisions - decisions that sometimes have severe consequences for those who are affected."

Consistency of Student Ratings

There is substantial agreement among researchers that student evaluations provide consistent feedback of general areas of an instructor's strengths and weaknesses, and can result in substantial improvement of specific target areas (e.g., Marsh and Roche, 1997, d'Apollonia and Abrami, 1997). Furthermore, Feldman (1988) found that students and faculty generally agree on what constitutes effective teaching and rank similarly its most important components.

Despite the generally positive reports on the consistency of student ratings, other accounts yield a less favorable picture. For example, Greenwald (1997) describes how he received the highest marks on a course, only to receive an appalling rating (lower by eight points on a scale of ten) on the same course the next year. He points out "the two juxtaposed ratings contained more than a mild hint that my students' responses were determined by something other than the (unchanged) course characteristics or the (presumably unchanged) instructor's teaching ability." Contradictory results are also reported by Follman (1983), who found in his study that when students were asked to name their best and worst teachers, 15% to 20% of the instructors appeared in both lists.

The Dr. Fox Effect

To demonstrate that a highly entertaining and expressive instructor can receive unduly high ratings, Naftulin et al. (1973) designed an experiment in which a professional actor introduced to the students as "Dr. Fox" gave a "highly expressive and enthusiastic lecture that was devoid of content, and received high overall ratings" (Watchel, 1998). This bias is commonly referred to as the Dr. Fox effect, and has been re-examined and fiercely contested in later studies (e.g.,

Abrami et al., 1982; Leventhal et al., 1976). The current consensus is that although the Dr. Fox effect does influence ratings, it is a relatively minor bias (Marsh, 1987).

Grading Leniency

One of the most researched issues in teaching evaluations is whether an instructor can increase his/her ratings by giving undeserved high grades to students. The topic is critical because if such bias is true, the use of student evaluations for administrative decisions is fundamentally undermined. Despite numerous studies on the effect of grading leniency, this issue is far from settled. Some researchers have found that a moderate correlation between grade leniency and ratings does exist (e.g., Chacko, 1983; Vasta and Sarmiento, 1979, Powell, 1978).

Opposite findings on grade leniency bias are reported by Greenwald and Gillmore (1997). In their article, a section is entitled "Yes, I can get higher ratings by giving higher grades." They present the results of their own research at the University of Washington, which included the collection of data over three or more semesters on university-wide samples of courses. Their conclusion was that for their sample, a grade increase from one standard deviation below the university mean to one standard deviation above the mean could increase an instructor's rating from half a deviation below the university mean rating to half a deviation above the mean university rating. In such case, using a normal statistical distribution, the instructor would get a boost from the 31st percentile to the 69th percentile.

Other Biasing Factors

Hewett et al. (1988) found that good grades on the first examination correlates positively to the ratings given to instructors, and that subsequent examinations have less effect on ratings. McKeachie (1979) and Gigliotti (1987) report that a very important biasing factor is the students' expectations about the instructor, i.e., the instructor's reputation affects his/her ratings. Feldman (1979) has found that the instructor's presence in the classroom tends to increase his/her rating, and that anonymous questionnaire responses tend to be more critical than those where the rater is identified. Rubin (1995) found that instructors with attractive physical appearance but authoritarian attitudes had less negative reviews of such authoritarian attitudes than similar instructors with less attractive physical appearance. Non-native speaker instructors with attractive physical appearance were less criticized for their accent than other non-native speaker instructors with less attractive physical appearance.

Feldman (1979) asserts that smaller class sizes lead to better ratings. He also concludes that elective courses usually have better reviews than required courses, as well as higher-level courses. Centra (1993) found that science and mathematics instructors receive lower rates than their liberal arts counterparts. Other findings include that the correlation between research activity and teaching effectiveness is near zero (i.e., one does not influence the other) (Centra, 1993), and that course difficulty correlates positively with ratings (Marsh, 1987).

Conclusion

This article has shown how teaching evaluations are intrinsically double-edged swords, one edge being their formative applications, and the other their summative applications. The mid-term evaluation example described here shows how student feedback can be formatively used to bring dramatic, immediate improvements to the classroom. Even considering the bias factors and other concerns discussed here, the majority of studies show that evaluations can provide insight into an instructor's basic strengths and shortcomings. If the formative edge is sharp, the other edge of the evaluation sword seems equally cutting. Summative applications, such as for salary adjustments, are repulsive to many faculty members (e.g., Newport, 1996). The author's anecdotal experience is that many instructors have faced a situation where a difficult assignment or an honest mistake while grading an exam disgruntles a couple of students, who then create a negative classroom environment. Common sense would indicate that the rating given to the instructor for such course would be lower than deserved. In fact, Jacobs (1987) found that 40% of surveyed students said that they have heard of "students plotting to get back at an instructor by collectively giving low ratings." It is hardly surprising that an instructor facing this scenario will not take seriously the students' feedback.

This article creates an intellectual baseline for further discussion and research geared to construction management education. What are the teaching evaluation practices used in construction management programs? How should they be shaped to the needs of construction management education? What are the perceptions about teaching evaluations from faculty, student and administrators? Uncritically accepting the results of existing analyses created for other fields of study would be shortsighted. It could be that the evaluation instruments used now have an unjust negative impact for construction management faculty and construction management education in general. It is important and urgent to find answers to these questions.

References

Abbott, R. D., Wulff, D. H., Nyquist, J. D., Rupp, V. A. and Hess, C. W. (1990) "Satisfaction with processes of collecting student opinions about instruction: the student perspective." *Journal of Educational Psychology,* 82, 201-206.

Abrami, P. C., Leventhal, L., and Perry, R. P. (1982) "Educational seduction." *Review of Educational Research,* 52, 446-464.

Baxter, E. P. (1991). "The TEVAL experience, 1983-88: the impact of a student evaluation of teaching scheme on university teachers." *Studies in Higher Education,* 16, 151-178.

Centra, J. A. (1993). *Reflective faculty evaluation.* San Francisco, CA: Jossey-Bass.

Chacko, T. J. (1983). "Student ratings of instruction: a function of grading standards." *Educational Research Quarterly,* 8, 19-25.

Christensen, K. and Rogers, L. (1992). "Teaching, service, and research in evaluation of construction management faculty for tenure and promotion." *Proceedings, 1992 Associated Schools of Construction Conference*, 79-84

Ciesielski, C. A. (1997). Tenure and promotion: a comparison between construction management and civil engineering." *Proceedings, 1997 Associated Schools of Construction Conference*, 21-32.

Clark, D. J., and J. Bekey (1979). "Use of small groups in instructional evaluation." POD Quarterly 1(2), 87-95

Cohen, P. A. (1980). "Using student ratings feedback for improving college instruction: a meta-analysis of findings." *Research in Higher Education*, 13, 321-341.

D'Apollonia, S., and Abrami, P. C. (1997). "Navigating student ratings of instruction." *American Psychologist*, 52 (11), 1198-1208.

Fabry, V. J., Eisenbach, R., Curry, R. R., and Golich, V. L. (1997). "Thank you for asking: classroom assessment techniques and students' perceptions of learning." *Journal on Excellence in College Teaching*, 8 (1), 3-21.

Feigenbaum, L. and Holland, N. (1997). "Using peer evaluations to assign grades on group projects." *Proceedings, 1997 Associated Schools of Construction Conference*, 75-80.

Feldman, K. A. (1979) "The significance of circumstances for college students' ratings of their teachers and courses." *Research in Higher Education, 10*, 149-172.

Feldman, K. A. (1988) "Effective college teaching from the students' and faculty's view: matched or mismatched priorities." *Research in Higher Education*, 28, 291-344.

Follman, J. (1983). "Student ratings of faculty teaching effectiveness: revisited." Paper presented at the annual meeting of the Association for the Study of Higher Education, Washington, D. C.

Gigliotti, R. J. (1987). "Are they getting what they expect?" *Teaching Sociology*, 15, 365-375.

Greenwald, A. G. and Gillmore, G. M. (1997). "Grading leniency is a removable contaminant of student ratings." *American Psychologist*, 52 (11), 1209-1217.

Greenwald, Anthony G. (1997). "Validity concerns and usefulness of student ratings of instruction." *American Psychologist*, 52 (11), 1182-1186.

Hauck, A. J. (1998). "Toward a taxonomy of learning outcomes for construction management education." *Proceedings, 1998 Associated Schools of Construction Conference*, 87-102.

Hewett, L., Chastain, G. and Thurber, S. (1988). "Course evaluations: are students' ratings dictated by first impressions?" Paper presented at the annual meeting of the Rocky Mountain Psychological Association, Snowbird, UT.

Jacobs, L. C. (1987). *University Faculty and Students' Opinions of Student Ratings.* Indiana Studies in Higher Education, #55 (Bloomington, IN, Bureau of Evaluation and Testing, Indiana University).

Leventhal, L., Abrami, P. C. and Perry, R. P. (1976). "Do teacher rating forms reveal as much about students as about teachers?" *Journal of Educational Psychology*, 68, 441-445.

Lin, Y. G., McKeachie, W. J., and Tucker, D. G. (1984). "The use of student ratings in promotion decisions." *Journal of Higher Education*, 55, 583-589.

Marsh, H. W. (1987). "Students' evaluation of university teaching: Research findings, methodological issues, and directions for future research." *International Journal of Educational Research, 11*, 253-388.

Marsh, H. W. and Dunkin, M. J. (1992) "Students' evaluations of university teaching: A multidimensional perspective." in: J. C. Smart (Ed.) *Higher Education: Handbook of Theory and Research, Vol.* 8.

Marsh, H. W., and Roche, L. A. (1997). "Making students' evaluations of teaching effectiveness effective." *American Psychologist*, 52 (11), 1187-1197.

Mckeachie, W. J. (1979). "Student ratings of faculty: a reprise." *Academe*, 65, 384-397.

McKeachie, W. J. (1997). "Student Ratings: the validity of use." *American Psychologist*, 52 (11), 1218-1225.

Murray, H.G. (1984). "The impact of formative and summative evaluation of teaching in North American Universities." *Assessment and Evaluation in Higher Education,* 9, 117-132.

Naftulin, D. H., Ware, J. E. and Donnelly, F. A. (1973). "The Doctor Fox lecture: A paradigm of educational seduction." *Journal of Medical Education,* 48, 630-635.

Newport, J. F. (1996). "Rating teaching in the USA: probing the qualifications of student raters and novice teachers." Assessment and Evaluation in Higher Education, 21 (1), 17-23.

Powell, R. (1978). "Faculty rating scale validity: the selling of a myth." *College English,* 39, 616-629.

Ramsden, P. and Dodds, A. (1989). *Improving Teaching and Courses: a guide to evaluation.* Parkeville, Melbourne: Centre for the Study of Higher Education, University of Melbourne.

Rubin, D. (1995). "Effects of language and race on undergraduates' perceptions of international instructors: further studies of language and attitude in higher education." Paper presented *at* the International Communication Association, Albuquerque, NM.

Ryan, J. J., Anderson, J. A. and Birchler, A. B. (1980). "Student evaluation: the faculty responds." *Research in Higher Education*, 12, 317-333

Segner, R. and Arlan, T. G. (1991). "Outcomes assessment in construction higher education." *Proceedings, 1991 Associated Schools of Construction Conference*, 49-52.

Simpson, R. D. (1995). "Uses and misuses of student evaluations of teaching effectiveness." *Innovative Higher Education*, 20 (1), 3-5.

Slobojan J. (1992). "Implementing outcome assessments for program accreditation." *Proceedings, 1992 Associated Schools of Construction Conference*, 29-34.

Spencer, P. A. and Flyr, M. L. (1992). "*The formal evaluation as an impetus to classroom change: myth or reality"?* Research/Technical Report. CA: Riverside.

Timpson, W. and Bendel-Simso, P. (1996). *Concepts and Choices for Teaching: meeting the challenges of Higher Education.* Madison, WI: Magna.

Vasta, R. and Sarmiento, R. F. (1979). "Liberal grading improves evaluations but not performance." *Journal of Educational Psychology*, 71, 207-211.

Wachtel, H. K. (1998). "Student evaluation of college teaching effectiveness: A brief review." *Assessment and Evaluation in Higher Education*, 23 (2), 191-211.

Wachtel, H. K. (1994). *A critique of existing practices for evaluating mathematics instruction*. Doctoral dissertation, University of Illinois at Chicago, *Dissertation Abstracts International*, 56.

Wulff, D. H., Staton-Spicer, A. Q., Hess, C. W. and Nyquist, J. D. (1985). "The student perspective on evaluating teaching effectiveness." *ACA Bulletin*, 53, 39-47.

Yoakum, B. (1994). "Program assessment – good management practice." *Proceedings, 1992 Associated Schools of Construction Conference*, 191-200